DOCUMENT RESUME

ED 441 808 TM 030 842

AUTHOR Pyo, Kyong Hyon
TITLE Assessing Dimensionality of a Set of Language Test Data with Simulated and Real Data Sets.
PUB DATE 2000-04-00
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Evaluation Methods; Factor Analysis; *Language Tests; Nonparametric Statistics; Simulation
IDENTIFIERS *Dimensionality (Tests)

ABSTRACT

The primary purpose of this study was to compare the performance of three procedures to assess dimensionality that were investigated by R. Nandakumar (1994) at different test conditions to reflect the characteristics of language test data. Procedures investigated were nonlinear factor analysis, the procedure of P. Holland and P. Rosenbaum, and W. Stout's nonparametric approach. To serve this purpose, simulated data consisting of 2,400 sets of unidimensional and multidimensional test data were used. The procedures were repeated with real language test data from the Test of English as a Foreign Language for 300, 500, and 1,000 examinees. Speculation of the real data structures was made on the basis of the findings from the simulation. Findings show that all three methodologies are able to discriminate between one- and two-dimensional test data. For simulated unidimensional test data, all three procedures were able to confirm unidimensionality, but for two-dimensional test data, the three procedures differed in their ability to detect the lack of unidimensionality. As statistics for assessing the degree of multidimensionality, both the approximate chi-square statistic and the DIMTEST T values were considered useful. An appendix contains six tables of study data. (Contains 6 tables and 36 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 808

Assessing Dimensionality of a Set of Language Test Data with Simuated and Real Data Sets

Kyong Hyon Pyo

University of Illinois at Urbana-Champaign

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

Kyong Hyon Pyo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

TM030842

BEST COPY AVAILABLE

Introduction

The majority of research studies that have applied IRT to problems with language testing data have not considered any checks of the unidimensionality assumption (e.g., Chen & Henning, 1985; Griffin, 1985; Pollitt & Hutchinson, 1987), while the results of some of the studies that have investigated either the assumption or dimensionality procedures do not provide a consistent picture (Jones, 1988; Choi & Bachman, 1992; Blais & Laurier, 1995). Davidson and Henning (1985), using exploratory factor analysis, found that two factors underlay scores from self-ratings, which they interpreted as multidimensional, while Henning, Hudson and Turner (1985), using Bejar's (1980) procedure, found scores from a multiple-choice EFL proficiency examination to be unidimensional. In Choi and Bachman (1992) and Blais and Laurier (1995) studies, they reported inconsistent results of different approaches, raised some concern about the uncertainty of the approaches to detecting dimensionality, and offered little solace to conscientious practitioners who want to be able to check the assumption before proceeding with IRT analyses. In those studies, they demonstrated that issues that have been widely recognized and discussed in the measurement literature are also problematic for language tests data: the dimensionality of a given test is difficult to assess, and is a matter of degree, so that the question is not whether a test is unidimensional or not, but rather the degree to which it depart from unidimensionality.

As for validating dimensionality procedures with simulated data sets, language testing is on its very early stage. Henning (1988) examined the effectiveness of Bejar's and several other approaches to detecting violations of unidimensionality with a set of data sets that were simulated to be bidimensional, and then assessed the consequences of

such violations for Rasch 1-parameter IRT model item and ability estimates. His results suggest that violation of unidimensionality does not affect Rasch estimates of item difficulty, but that Rasch person ability estimates are seriously affected. What this research indicates that language tests may be multidimensional and this may seriously affect at least the ability estimates obtained from at least one IRT model – the Rasch model.

In general measurement fields, researchers have constantly and rigorously proposed dozens of descriptive and inferential statistics to assess dimensionality, or more commonly, departure from the assumption of unidimensionality. Many simulation studies have been conducted to see how well a dominant ability can be recovered with different approaches. Hattie (1985), Hambleton and Rovinelli (1986), and Berger and Knol (1990) have reviewed several procedures for assessing dimensionality and found none of them satisfactory for different conditions of tests.

Starting with linear factor analysis used to assess the dimensionality of the latent space underlying the responses to a set of items, researchers empirically tested its dependability as a dimensionality check and reported for test developers and users. A number of methodological problems associated with using linear factor analyses to assess dimensionality were recognized. For example, factor structure of item responses depends on characteristics of item parameters such as difficulty levels of items and guessing levels of multiple-choice items (Hulin, Drasgow, & Parsons, 1983; Zwick, 1987). Furthermore, the linear factor model has a tendency of overestimating the number of underlying dimensions in the data. Consequently, new methods have constantly been sought to assess dimensionality has been witnessed for the last decade. Then, it entailed a series of

simulation studies to ensure their quality for dimensionality methods. Some of the recently developed methods include nonlinear factor analysis (McDonald & Ahlawat, 1974); Bejar's procedure (Bejar,1980); residual analysis (Hambleton & Swaminathan, 1985); Bock's full information factor analysis (Bock, Gibbons, & Muraki, 1985); Holland and Rowenbaum's test of unidimensionality, monotonicity, and conditional independence (Halland & Rosenbaum, 1986); Roznowski, Tucker, and Humphreys' procedures (1991); and Stout's unidimensionality procedure DIMTEST (Stout, 1987), just to name a few.

The main purpose of this study is to bring into attention the needs for language testers to be selective of methods of dimensionality by testing their credibility when deciding which methods are appropriate to use for their data. Considering the distinctiveness of language test data in terms of test characteristics such as extremely high correlations between latent traits, when assumed to be multidimensional, we cannot assume a proper application of methods to language test data, simply because those methods have been empirically validated in many other areas.

To serve the purpose of the study, three procedures (Nonlinear factor analysis (NLFA), Holland and Rosenbaum's procedure (H & R), and Stout's nonparametric approach) to assess dimensionality that were investigated by Nandakumar (1994) were compared among themselves and across these two studies (Nandakumar, 1994 and mine). The rationale of a replication study with the same procedures is twofold: 1) language testing-specific data need to be explicitly under focus, 2) more simulation process needs to be done beyond data sets produced only once under different conditions. Two thousand four hundred sets of unidimensional and multidimensional test data were

simulated and used to study the performance of all three procedures for assessing dimensionality. The same procedures were then repeated with real language test data.

Description of Procedures

Nonlinear Factor Analysis

McDonald (1967) and Takane and De Leeuw (1987) have shown that common IRT functions are a special case of a more general nonlinear factor analytic model, and that they are mathematically equivalent. Based on this IRT-NLFA relationship, some researchers have suggested that the assessment of dimensionality should be based on the analysis of the residual covariance matrix obtained after fitting an *m*- factor NLFA model to an item response matrix, where *m* is equal to the number of factors hypothesized to underlie the items response matrix (Gessaroli, 1994; Goldstein, 1980; Goldstein & Wood, 1989; McDonald, 1981, 1994). These NLFA models can be used to assess not only departure from unidimensionality but also the fit of more complex models, for example, a two- or three-dimensional structure to a matrix of item responses.

Unidimensionality of the latent ability space would theoretically imply zero residual covariances between all pairs of binary item responses, U_j and U_k, at fixed ability levels; that is,

$$\mathrm{cov}(U_j, U_k \mid \theta) = 0, \; j \neq k \qquad (1)$$

McDonald (1981) refers to equation (1) as the weak principle of local independence. Several researchers have stated that McDonald's model and accompanying weak principle of local independence appear to be promising with regard

to the assessment of dimensionality (Goldstein & Wood, 1989; Hambleton & Rovinelli, 1986; Hattie, 1984).

An approximate chi-square statistic, based on McDonald's NLFA model, was investigated by De Champlain (1992) and Gessaroli and De Champlain (1996) as a potentially useful procedure for assessing dimensionality. The procedure proposed for assessing the dimensionality of a set of items is based on the computation of an approximate chi-square test. This statistic involves estimating parameters for an m-factor model using the NLFA approach outlined by McDonald (1967) and implemented by Fraser and McDonald (1988) in the computer program NOHARM. This approximate chi-square statistic is defined as

$$\chi^2 = (N-3)\sum_{i=1}^{p}\sum_{j=1}^{i-1} Z_{ij}^{2(r)} \qquad (2)$$

where $z_{ij}^{2(r)}$ is the square of the Fisher z corresponding to the residual correlation between items i and j, ($i, j = 1 \ldots p$) and N is the number of examinees in the sample. This statistic is approximately distributed as a central chi-square with $df = p(p-1)/2-t$, where p is equal to the number of items and t is the total number of independent parameters estimated in the model. The specific computational steps for this approximate χ^2 statistic are given in Gessaroli and De Champlain (1996). With simulated unidimensional data sets, the empirical Type I error rates tended to be lower than the normal α for the shorter test lengths examined but increased to values close to expected α probabilities for the longer tests. Also the empirical Type I error rates obtained for the approximate χ^2 statistic were not affected by the nonnormality of the latent trait distributions (De Champlain & Tang, 1993). With simulated multidimensional data sets, rejection rates based on the

approximate χ^2 statistic were generally high, even in some instances where data sets contained as few as 15 items and 500 examinees, which was not the case for Stout's T statistic (De Champlain, 1992; Gessaroli & De Champlain, 1996).

Holland and Rosenbaum's Test of Lack of Fit of a Unidimensional, Monotone, and Conditional Independence Model

Rosenbaum (1984) and Holland and Rosenbaum (1986) have proved theorems concerning conditional association that can be applied to assess dimensionality. The basic notion in Holland and Rosenbaum's (H & R) theorems is that if the items are locally independent, unidimensional, and the item characteristic curves are monotone, then the items are conditionally positively associated. Specifically, the conditional covariances between any pair of item response functions of a set of unidimensional dichotomous item responses given any function of the remaining items responses will be nonnegative. The test of this relationship can be specified as

$$H_0 : \mathrm{Cov}\left(X_i, X_j \mid \sum_{i,j \neq k} X_k\right) \geq 0 \quad \text{versus } H_1 : \mathrm{Cov}\left(X_i, X_j \mid \sum_{i,j \neq k} X_k\right) < 0 \qquad (3)$$

Conditional association for each pair of items are tested, given the number-right score on the remaining items. The Mantel-Haenszel test (M-H) is used to test this hypothesis. To perform the M-H test on a given pair of items, a 2×2 contingency table is constructed for the pair for each of the possible number-right socres on the remaining items. The cell values of a 2×2 table for item pair i and j for examinees with total score k (k = 1,2,..., K) on the remaining items can be denoted as the following: the number of examinees who got both item i and item j correct (n_{11k}), the number of examinees who got both item i and item j incorrect (n_{00k}), the number of examinees who got item i correct and item j

BEST COPY AVAILABLE

incorrect (n_{10k}), and the number of examinees who got item i incorrect and item j correct (n_{01k}). The M-H statistic is then given by

$$Z = \frac{n_{11+} - E(n_{11+}) + 1/2}{\sqrt{V(n_{11+})}}, \tag{4}$$

$n_{11+} = \sum_{k=1}^{K} n_{11k}$, and $E(n_{11k})$ and $V(n_{11+})$ are the expectation and the variance of n_{11+} given by

where

$$E(n_{11k}) = \sum_{k=1}^{K} \frac{n_{1+k} n_{+1k}}{n_{++k}} \tag{5}$$

and

$$V(n_{11+}) = \sum \frac{n_{1+k} n_{0+k} n_{+1k} n_{+0k}}{n_{++k}^2 (n_{++k} - 1)}. \tag{6}$$

The plus subscript in Equation 5 and 6 denotes the summation over that subscript. The computed Z value is compared to the lower tail of the standard normal distribution. A statistically significant Z implies that the pair of items in question is not conditionally associated, given the sum of the remaining items, and is thus inconsistent with the unidimensional model. In this manner, the M-H statistic is computed for all $N(N-1)/2$ pairs of items, where N is the total number of items in a test. If a large number of pairs are shown not to be conditionally associated, then the unidimensional assumption is inappropriate.

Ben-Simon and Cohen (1990) have demonstrated the application of H & R's approach to assess dimensionality. Ben-Simon and Cohen found the H & R approach was conservative and erroneously misclassified nearly half of the multidimensional item pools they analyzed as unidimensional. Zwick (1987) found H & R's approach to be

consistent with other procedures investigated in assessing unidimensionality of NAEP reading data.

DIMTEST

Stout proposed a nonparametric procedure that is based on his notions of essential independence and essential dimensionality (Nandakumar, 1991; Nandakumar & Stout, 1993; Stout, 1987, 1990). Essential dimensionality corresponds to the number of latent traits that are required to satisfy the assumption of essential independence given by

$$\frac{1}{n(n-1)} \sum_{1 \le i \le j \le n} |\mathrm{Cov}(U_i, U_j | \theta)| \to 0, n \to \infty, \tag{7}$$

that is, a mean absolute residual covariance value that tends towards zero at fixed levels of the latent trait as the number of items increases towards infinity. Stout (1987, 1990) and Nandakumar and Stout (1993) proposed several versions of the T statistic to test the assumption of essential unidimensionality (d_e), that is,

$$\begin{aligned} &\mathrm{H}_0 : d_e = 1, \\ &\mathrm{H}_1 : d_e > 1, \end{aligned} \tag{8}$$

where d_e corresponds to the number of dimensions required to satisfy the assumption of essential independence. Briefly, the procedure entails dividing a set of items into two subtests labeled AT1 and AT2, as well as a partitioning test, or PT. The AT1 items are selected as the unidimensional subset, usually following a linear factor analysis of a tetrachoric correlation matrix. The AT2 items are chosen to have an item difficulty distribution as similar as possible to that of the AT1 items. Stout (1987) and Nandakumar and Stout (1993) point out that the T statistic value is often inflated due to the presence of guessing and a difficulty factor following the linear factor analysis

performed to select AT1 items. The purpose of the AT2 subtest is to correct for these statistical biases. The T_2 statistic was selected for the present investigation because of its demonstrated low Type I error and high power (Nandakumar & Stout, 1993). The T_2 statistic can be defined as follows:

$$T_2 = \frac{T_{L,2} - T_b}{\sqrt{2}}, \tag{9}$$

where

$$T_{L,2} = \frac{1}{K^{1,2}} \left(\sum_{k=1}^{k} \frac{X_k}{S_k} \right), \tag{10}$$

K = the number of subgroups based on the PT items subscore, k_i = the ith subgroup of examinees based on the PT item subscore, and S_k = the standard error of the T_2 statistic within the kth subgroup. The $T_{L,2}$ and T_b statistics are identical with the exception that the former is computed for AT1 items, and the latter for AT2 items. The T_2 statistic is asymptotically distributed as a standard normal variate under the null hypothesis of unidimensionality. In a series of studies carried out by Nandakumar and Stout (1993), the T_2 statistic was found to be quite accurate in the case of multidimensional data sets, except when the test contained few items (less than 25) and the sample size were small (less than 750 examinees).

Purpose

1. To examine the empirical Type I error rates calculated for the approximate chi-square statistic, Holland and Rosenbaum's Test (H & R), and Stout's T statistic with unidimensional data sets simulated with various test lengths and sample sizes.

2. To examine the rejection rates obtained for the approximate chi-square statistic, Holland and Rosenbaum's Test (H & R), and Stout's T statistic with two-dimensional item response matrices generated to vary as a function of sample size, test length, and degree of correlation between the latent traits.
3. To examine the capability of the approximate chi-square statistic, Holland and Rosenbaum's Test (H & R), and Stout's T statistic with three language test data sets.
4. To compare the three procedures across two studies (Nandakumar, 1994 and mine).

Method

Unidimensional Data Set Simulations

Dichotomous unidimensional item responses were simulated according to the three-parameter logistic model given by:

$$P_i(x_i = 1 \mid a_i, b_i, c_i, \theta_j) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}}. \qquad (11)$$

100 data sets were generated to vary according to two different test lengths (25 and 50 items) as well as three sample sizes (300, 500, 1000 examinees). The 25-item tests were used double to simulate 50-item data sets; that is, the item parameters for items 26 to 50 were identical to those for items 1 to 25. To simulate item responses that are typical of those with the paper-and-pencil Test of English as Foreign Language (TOEFL), 25 IRT parameters were selected from a listening section of TOEFL and used in the item response generation process. The item parameters that were chosen to simulate unidimensional item responses are shown in Table 1.

Table 1 goes here

Latent trait values were simulated according to a $N(0,1)$ distribution. 100 times of replication made a total of 600 unidimensional data sets by each cell of this 2 (test length) x 3 (sample size) design. The fit of a unidimensional model was then tested for each of the 600 unidimensional data sets using NOHARM, H & R, and DIMTEST with unidimensional specification. The approximate chi-square statistic was computed for each data set using computer program CHIDIM (De Champlain & Tang, 1997).

Two-Dimensional Data Set Simulations

Dichotomous two-dimensional item responses were simulated based on the two-parameter compensatory multidimensional model given by:

$$P(i=1 \mid \theta) = c_i + (1-c_i)\frac{e^{1.7(a_1\theta_1 + a_2\theta_2 + d_i)}}{1+e^{1.7(a_1\theta_1 + a_2\theta_2 + d_i)}} \qquad (12)$$

All two-dimensional data sets were generated to reflect the unidimensional conditions with a function of the same two test lengths (25 and 50 items) and three sample sizes (300, 500, 1000 examinees). The 50-item data sets were composed of two 25-item tests as done for unidimensional data sets. As for the data structure, each item was related to only one dimension. Past research has shown that the correlations between factors examined in the dimensionality study in language testing is very high, but not with a

BEST COPY AVAILABLE

complex data structure. The correlation between the two latent traits was therefore set at 0.5, 0.65, and 0.8 in the two-dimensional simulation. In addition, the dimension dominance was fixed to one level, that is, 40:60, 40 % of the items indicative only of Dimension 1 (θ_1), and the remaining 60% reflective only of Dimension 2 (θ_2). This dimension dominance structure accords that of real data sets to be investigated in the later section. Finally, 100 times of replication simulated a total of 1800 two-dimensional data sets by each cell of this 2 (test length) x 3 (sample size) x 3 (latent trait correlation) design.

Since none of the language tests were available as having true two-dimensional parameters, the item parameters obtained from Law Scholastic Aptitude Test (LSAT) data shown in table 2 were utilized for the two-dimensional data simulation. The power of the three statistics in correctly rejecting the assumption of unidimensionality with two-dimensional simulated test structures was under main focus. Unlike the case with unidimensional data sets, both unidimensional and multidimensional models were fitted to each two-dimensional data set.

Table 2 goes here.

The Real Test Data in Language Testing

The real test data used in this study were selected from three different sources. First Certificate in English (FCE) and Test of English as a Foreign Language (TOEFL) data were obtained from the Comparability of Tests in Cambridge Study (CTCS) (1990). Then, the English Placement Test (UIUC-EPT) data were gathered from University of Illinois (1999). All the data sets contain two language skills as dimensions (listening and

reading). For all data sets, examinees with one or more missing items were deleted from the analyses. Test lengths, sample sizes, and number of items of each trait for all real tests are given in Table 3.

Table 3 goes here

Results

Unidimensional Data Set Analyses

The number of false rejections of the assumption of unidimensionality for the 100 unidimensional data sets simulated in each condition is summarized in Table 4.

Approximate chi-square statistic empirical Type I error rates (NOHARM)

The empirical Type I error rates tended to be below or near the alpha level (.05). In fact, the maximum number of rejections of the assumption of unidimensionality in any given condition was 5 out of 100 (.05) for data sets simulated to contain 50 items and 1,000 examinees. Logit-linear analysis results show that the frequency of rejections of the assumption of unidimensionality is not explained by any of the independent variables entered.

There seemed a steady increase in the rejection rate as sample size becomes larger across two test lengths. There were two false rejections (.01) of the assumption of unidimensionality for data sets simulated to include 300 examinees as opposed to nine incorrect rejections (.045) for data sets with 1000 examinees.

H & R Procedure

The empirical type I error rates obtained with the H&R approach were controlled below α ($\alpha = .05$) level all across conditions, strong supporting the unidimensional nature of these data sets.

DIMTEST

The number of incorrect rejections of the assumption of unidimensionality, based on the DIMTEST was slightly larger for data sets with 50 items than that found with NOHARM and H & R Procedure. Empirical Type I error rates ranged from 0.00 (for data sets that included 25 items and 300, 500 examinees) to 0.09 (for data sets that included 50 items and 500 examinees). These results show a controlled Type I error rate precision when using the DIMTEST to determine whether an item response matrix was unidimensional. The results obtained from the logit-linear analysis indicate that neither test length nor sample size is required to adequately explain the frequencies of rejection rates.

Table 4 goes here.

Multidimensional Data Set Analyses

Approximate chi-square statistic empirical Type I error rates (NOHARM)

The number of rejections of the assumption of unidimensionality obtained using the approximate chi-square statistic is shown in Table 5 for the simulated two-dimensional data sets.

Results clearly show that the approximate chi-square statistic was able to consistently identify the (true) multidimensional nature of the simulated data sets. The assumption of unidimensionality was correctly rejected for all simulated data sets (1,800) 100 %. The logit-linear analysis results indicate that a model including only the dependent variable rejection decision was sufficient to explain the observed frequencies. These results suggest that NOHARM has great power for rejecting unidimensionality.

H & R Procedure

There was a considerably greater degree of variability in rejection rates based on the H & R analysis. Rejection rates ranged from 2 out of 100 (25-item data sets simulated to contain both 300 and 500 examinees with .80 correlation between latent traits) to 44 out of 100 (25 item data sets for 500 and 1,000 examinees with .50 correlation). Logit-linear analysis results yield a model that includes the main effect of latent trait correlation, L^2 (1) = 39.550, p = 1.000.

With respect to the main effect of latent trait correlation, results indicate that the number of false acceptances of unidimensionality was significantly greater for data sets in which a correlation of 0.65 was specified between latent traits (544 out of 600 or 0.91), and a correlation of 0.80 (579 out of 600 or 0.97) than data sets with a correlation of 0.50 (245 out of 600 or 0.41). That is, the number of false acceptances of the assumption of unidimensionality increased sharply for 0.65 and 0.80 correlation data sets from a

rejection of 0.41 that includes a correlation of 0.50 between traits to 0.91 and 0.97. In overall, there were high rates of false acceptances of unidimensionality for data sets to simulate two-dimensional responses across all conditions.

DIMTEST

Similar to the results of H & R procedure, DIMTEST analysis produced a considerably a great degree of variability in rejection rates. Rejection rates ranged from 4 out of 100 (data sets simulated to contain 25 items, 300 examinees, and a correlation of 0.80 between latent traits) to 100 out of 100 (some of conditions with 50 items). Logit-linear analysis results showed a model that included the main effects of test length, sample size as well as latent trait correlation, L^2 (18) = 12.01, p = 0.95.

With respect to the main effect of test length, results indicates that the number of failures to reject unidimensionality was significantly greater for 25 item data sets (630 out of 900 or 0.7) than for 50 item data sets (243 out of 900 or 0.27). Regarding the effect of sample size, the number of false acceptances of unidimensionality dropped stably as the number of sample size increased: 523 out of 600 (0.87) for the 300 examinee data sets, 264 out of 600 (0.44) for 500 examinees, and 155 out of 600 (0.26) for 1,000 examinees. Finally, findings obtained with regard to the effect of the latent trait correlation yield that the number of false acceptances of unidimensionality was getting greater for data sets with greater correlations: 257 out of 600 (0.43) for a correlation of 0.50, 302 out of 600 (0.50) for 0.65, and 378 out of 600 (0.63) for 0.80.

In summary, all three procedures—NLFA, H & R approach, and DIMTEST—were all good in supporting unidimensionality. However, they showed differential power

in multidimensional data sets. NLFA demonstrated the greatest power in detecting multidimensionality, rejecting correctly the assumption of unidimensionality for all data sets of varying conditions. The H & R approach demonstrated a severe lack of power across all the condition, which means the H & R is very conservative for detecting multidimensionality. Next, the rejection rates indicate that DIMTEST demonstrates varying power depending on different conditions (e.g., sample size, number of items, and trait correlation are all contributing elements to the power of DIMTEST).

Table 5 goes here.

The Real Test Data

The results of NOHARM, H & R procedure, and DIMTEST for real data sets are presented in Table 6 and Table 7. Based on χ^2 statistics, a one-factor model was rejected for all three data sets, which means those data sets are not unidimensional. Subsequently, a two-factor model was implemented to see if the data sets have a two-dimensional structure.

The goodness-of-fit statistics for the two-factor model indicate that FCE and UEPT data sets fit well, but not TOEFL data. Since TOEFL data was rejected for both one-dimensional and two-dimensional assumptions, it is assumed that a higher order of factor model was needed with approximate chi-square statistics. As a next step to investigating the TOEFL data structure, homogeneous items split based on the test section (e.g., vocabulary and reading items) were separately tested for a one-factor NLFA

model. The results of a one-factor NLFA are as follows: for the vocabulary subtest, χ^2 (377) = 899, $p < .000$, and for the reading subtest, χ^2 (377) = 730, $p < .000$. The results confirm those from the two-factor model in that both of subtests, vocabulary and reading test, reject the unidimensional assumption

---

Table 6 goes here.

The results based on the H & R test indicate that unidimensionality is not expected to be rejected for all three data sets, which is not surprising, considering the credibility that was found from data simulations. However, unlike the extremely low power of the H & R test based on the simulation data sets, the H & R test rejected the unidimensionality assumption for FCE and TOEFL. For data set UIUC-EPT, the number of significant negative partial correlations is well below the nominal level $t\alpha$, and retains the unidimensional assumption. On further examination of the H & R results, it was found that the M-H Z values for many of the item pairs, where they were assumed to be measuring different traits, did not reach significance level. The H & R approach produces consistent results as NLFA for the other data sets, FCE and TOEFL, which confirms the nature of multidimensionality in the data structure.

The results of DIMTEST under the hypothesis of essential unidimensionality are very similar to those of NFLA and the H & R Test. The only difference from the results of NLFA is that, for FCE, the T value associated with DIMTEST indicate that the data can be approximated by an essential unidimensionality model, but the decision is not clear because the T value of DIMTEST is in the borderline region.

Table 7 goes here.

Discussion

The primary purpose of this study was to compare the performance of three procedures to assess dimensionality that were investigated by Nandakumar (1994) at different test conditions to reflect the characteristics of language test data. To serve this purpose, simulation data were utilized. Furthermore, speculation of the real data structures was made on the basis of the findings obtained from simulation data. Based on this limited study, findings demonstrated that all three methodologies were able to discriminate between one-and two-dimensional test data. For simulated unidimensional test data, all three procedures were able to confirm unidimensionality. For two-dimensional test data, however, the three procedures differed in their ability to detect the lack of unidimensionality.

As opposed to the findings that Nandakumar (1994) reported on her study that DIMTEST showed the highest power in rejecting the null hypothesis of essential unidimensionality for all two-dimensional data sets and that the nonlinear factor analysis and the H & R approach had a modest lack of power for two-dimensional data sets, the approximate χ^2 statistic performed extremely well in identifying two-dimensional models with the accuracy of the statistic 100% (see Table 5). This is indicative of good power, based on the conditions and data simulated in this study. However, DIMTEST showed a wide range of accuracy from 7% to 100%. In line with many studies (De Champlain and Gessaroli, 1991, 1996), DIMTEST suffered from power decrease in varying conditions, specifically for the test length of 25. The results of the logit-linear analysis indicated

BEST COPY AVAILABLE

better accuracy of the statistic with increases in test length, sample size, and correlation between latent traits.

The results of the H & R approach based on simulation data in the present study were consistent with those from Nandakumar (1994). The accuracy of the H & R test in correctly rejecting unidimensionality was never higher than 44% all across the test conditions. This study supports the findings of the Ben-Simon and Cohen (1990) and Zwick (1987) studies that the H & R is too conservative to be used as a dimensionality tool.

For real data sets, the results of nonlinear factor analyses were in overall consistent with the findings of DIMTEST. The two analyses rejected the assumption of unidimensionality for TOEFL and UIUC-EPT data sets. The DIMTEST T value of 1.29 for FCE data achieved a minimal significance level ($p = .09$) in testing the assumption of an essential unidimensionality. Based on the great dependability of the approaches found from simulation studies, all of the three real data sets were determined to be multidimensional (at least bi-dimensional).

In summary, as statistics for assessing the degree of multidimensionality, both the approximate chi-square statistic and the DIMTEST T values are considered useful. Problems indicated by Nandakumar (1994) that determining the appropriate number of polynomial terms to retain in the model is arbitrary for nonlinear factor analysis was solved by the approximate chi-square statistic that aids in assessing the fit of nonlinear models. For problems recognized with DIMTEST that it loses power when both the test length and the sample is small, the results in my study showed support the characteristic of the statistic in that the T statistic did not perform well in conditions with a test length

of 25 items and a sample size of 300. Rejection rates with true two-dimensional data sets of DIMTEST were generally acceptable for a test length of 50 items except for one condition of a sample size of 300 (varying from 30 out of 100 to 14 out of 100 data sets), which indicates that a sample size should be greater than 500 for a test with a large set of items.

A great deal of emphasis should be made on the characteristics and weaknesses of different approaches to assessing dimensionality. Various approaches lead to different and sometimes conflicting results. There has been a misleading practice in language testing that linguistic and psychological concepts were defined only through a specific approach based on researchers' preference. Since dimensionality of a language test as part of the construct validity study is never fully defined due to its complexity, the issue of which approach is better one should be raised and the validity of different approaches should be evaluated with caution in language testing context.

References

Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-96.

Ben-Simon, A., Cohen, Y. (1990, April). Rosenbaum's test of unidimensionality: Sensitivity analysis. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Berger, M. P., & Knol, D. L. (1990, April). On the assessment of dimensionality in multidimensional item response theory models. Paper presented at the meeting of the American Educational Research Association, Boston.

Blais, J-G., & Laurier, M. (1995). The dimensionality of a placement test from several analytical perspectives. Language Testing, 12, 72-97.

Bock, D. R., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. Applied Psychological Measurement, 12, 261-280.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. Language Testing, 2, 155-63.

Choi, I-C., & Bachman, L. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. Language Testing, 9, 51-78.

De Champlain, A. & Tang, K. L. (1993, April). The effect of nonnormal ability distribution on the assessment of dimensionality. Paper presented at the meeting of the National Council on Measurement in Education, Atlanta, GA.

De Champlain, A. (1992). Assessing test dimensionality using two approximate chi-squaure statistics. Unpublished doctoral dissertation, University of Ottawa, Ottawa, Canada.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. Multivariate Behavioral Research, 23, 267-269.

Gessaroli, M. E., & De Champlain, A. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the response to a set of items. Journal of Educational Measurement, 33, 157-179.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology, 33, 234-246.

Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. British Journal of Mathematical and Statistical Psychology, 42, 139-167.

Griffin, P. (1985). The use of latent trait models in the calibration of tests of spoken language in large-scale selection-placement programs. In Y. P. Lee, A. C. Fok, R. Lord, & G. Low (Eds.), New directions in language testing (pp. 149-61). Oxford: Pergamon Press.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and application. Boston: Kluwer.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of test and items. Applied Psychological Measurement, 9, 139-64.

Henning, G. (1988). The influence of test and sample dimensionality. Language Testing, 5, 83-99.

Henning, G. (1989). Meanings and implications of the principle of local independence. Language Testing, 5, 83-99.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. The Annals of Statistics, 14, 1523-1543.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory. Homewood, IL: Dow Jones-Irwin.

Jones, P. B. (1988). Assessment of dimensionality in dichotomous-scored data, using multidimensional scaling: Analysis of HSMB data. (ERIC Document Reproduction Service No. ED 295979)

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 27, 82-89.

McDonald, R. P. (1967). Nonlinear factor analysis. Psychometrika Monograph, 32 (15, Pt. 2)

McDonald, R. P. (1994). Testing for approximate dimensionality. In D. B. D. Laveault Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), Modern theories of measurement: problems and issues (pp. 63-86). Ottawa, Ontario: Edumetrics Research Group, University of Ottawa.

Nandakumar, R. & Stout, W. F. (1993). Refinement of Stout's procedure for assessing latent trait dimensionality. Journal of Educational Statistics, 31, 17-35.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. Journal of Educational Measurement, 28, 99-117.

Nandakumar, R. (1994). Assessing the dimensionality of a set of item responses—Comparison of different approaches. Journal of Educational Measurement, 31, 17-35.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessment: Rasch partial credit analysis of performance in writing. Language Testing, 4, 71-92.

Rosenbaum, P. (1984). Testing the local independence assumption in item response theory (Tech. Rep. No. 84-85). Princeton, NJ: Educational Testing Service.

Roznowski, M., Tucker, L. R., & Humphreys, L. G. (1991). Three approaches to determining the dimensionality of binary items. Applied Psychological Measurement, 15, 109-127.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. Psychometrika, 52, 393-408.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. Journal of Educational Measurement, 24, 293-308.

Appendix

Table 1

True Unidimensional Item Parameters Based on TOEFL Data

Item	a	b	c
1	0.665	-0.410	0.237
2	0.626	-1.185	0.330
3	0.461	-0.231	0.282
4	0.941	-1.606	0.366
5	0.650	-1.142	0.293
6	0.903	-1.687	0.202
7	0.361	0.887	0.215
8	2.820	-1.172	0.387
9	0.446	0.967	0.204
10	4.690	2.645	0.251
11	0.903	-1.687	0.202
12	0.361	0.887	0.215
13	2.820	-1.172	0.387
14	0.446	0.967	0.204
15	4.690	2.645	0.251
16	0.203	0.181	0.258
17	0.789	-1.751	0.428
18	0.470	0.292	0.210
19	0.458	1.274	0.207
20	0.322	0.212	0.213
21	0.602	0.638	0.200
22	0.387	0.916	0.210
23	0.427	0.976	0.206
24	0.537	0.441	0.193
25	0.373	0.500	0.202

Table 2

True Two-dimensional Item Parameters Based on LSAT Data

Item	a1	a2	MDIF	c
1	0.62	0	-1.71	0.12
2	0.81	0	0.16	0.16
3	0.84	0	0.08	0.14
4	0.88	0	0.85	0.18
5	0.91	0	1.37	0.24
6	0.78	0	0.26	0.21
7	0.70	0	-1.36	0.12
8	0.54	0	-0.61	0.12
9	0.79	0	1.14	0.16
10	0.77	0	0.28	0.25
11	0	0.65	-0.89	0.12
12	0	0.88	1.08	0.06
13	0	0.96	0.92	0.19
14	0	0.95	1.52	0.22
15	0	0.80	-1.14	0.12
16	0	0.75	-0.32	0.12
17	0	0.55	0.63	0.12
18	0	0.68	0.55	0.12
19	0	0.80	-0.09	0.15
20	0	0.57	0.26	0.21
21	0	0.76	-0.15	0.20
22	0	0.91	1.21	0.15
23	0	0.88	1.08	0.09
24	0	0.87	0.94	0.13
25	0	0.90	1.02	0.21

Table 3

Description of real data sets

Test	Length	Sample size	Number of items of each trait	
			Trait 1	Trait 2
FCE	40	935	15	25
TOEFL	58	1247	29	29
EPT	27	316	12	15

Table 4

Rejections of Unidimensionality Per 100 Trials for Unidimensional Data Sets

	Fit Statistic					
	Approximate Chi-Square (NOHARM)		H & R test		Stout T (DIMTEST)	
Sample Size	25 Items	50 Items	25 Items	50 Items	25 Items	50 Items
300	0	2	0	0	0	8
500	1	2	2	1	0	9
1000	4	5	2	0	5	2

Table 5

Rejections of Unidimensionality Per 100 trials for Two-Dimensional Data Sets:

		Fit Statistic					
		Approximate Chi-Square (NOHARM)		H & R test		Stout T (DIMTEST)	
Trait Correlation	Sample Size	25 Items	50 Items	25 Items	50 Items	25 Items	50 Items
$r\theta_1\theta_2 = 0.50$	300	100	100	42	39	15	30
	500	100	100	44	40	34	100
	1000	100	100	44	36	74	100
$r\theta_1\theta_2 = 0.65$	300	100	100	8	11	10	19
	500	100	100	10	10	21	96
	1000	100	100	8	9	52	100
$r\theta_1\theta_2 = 0.80$	300	100	100	2	4	4	14
	500	100	100	2	6	7	78
	1000	100	100	3	4	26	93

Table 6

Results of Analyses of NOHARM (1- & 2- Factor) with Real Data Sets

	NOHARM (One-Factor) $H_0: \text{cov}(U_j, U_k \mid \theta) = 0, j \neq k$			NOHARM (Two-Factor) $H_0: \text{cov}(U_j, U_k \mid \theta) = 0, j \neq k$		
Name	χ^2	p<	Decision based on NOHARM	χ^2	p<	Decision based on NOHARM
FCE	1524.13	.00	Reject H_0	977.00	.19	Accept H_0
TOEFL	3617.94	.00	Reject	2453.19	.00	Reject
UIUC-EPT	463.38	.00	Reject	350.76	.14	Accept

Table 7

Results of Analyses of H & R Test and DIMTEST with Real Data Sets

	H & R Test $H_0: \text{Cov}(X_i, X_j \mid \sum_{i,j \neq k} X_k) \geq 0$			DIMTEST $H_0: d_L = 1$		
Name	No. of item pairs	*No. of pairs significant at level α	Decision based on H & R Test	T	p<	Decision based on DIMTEST
FCE	780	68	Reject H_0	1.29	.09	Accept H_0
TOEFL	1653	113	Reject	5.23	.00	Reject
UIUC-EPT	351	7	Accept	2.89	.00	Reject

* significant at .05 level